Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models

Terra Blevins[1] Luke Zettlemoyer[1,2]
[1] University of Washington [2] Meta AI

English pretrained models transfer across languages!

How does this happen?

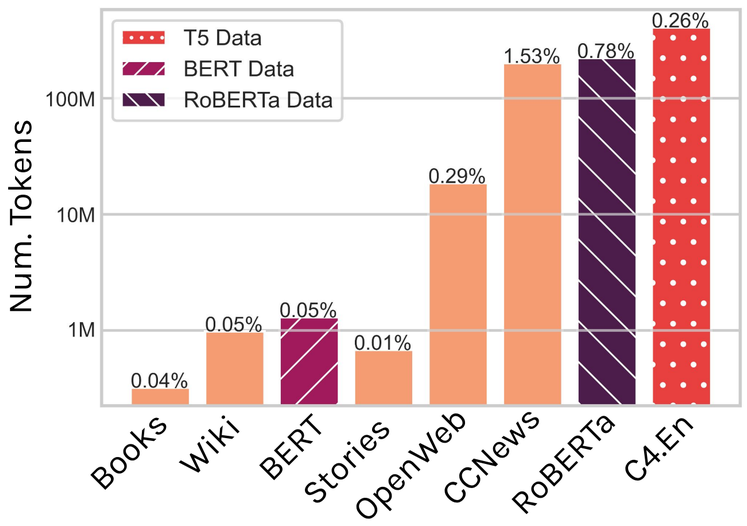
We examine this phenomenon with 3 experiments:

(1) Estimating Language Leakage with automatic language classifier + manual analysis
(2) Measuring Cross-Lingual Transfer
(3) Correlation Study of Factors Related to Cross-Lingual Transfer

Findings:

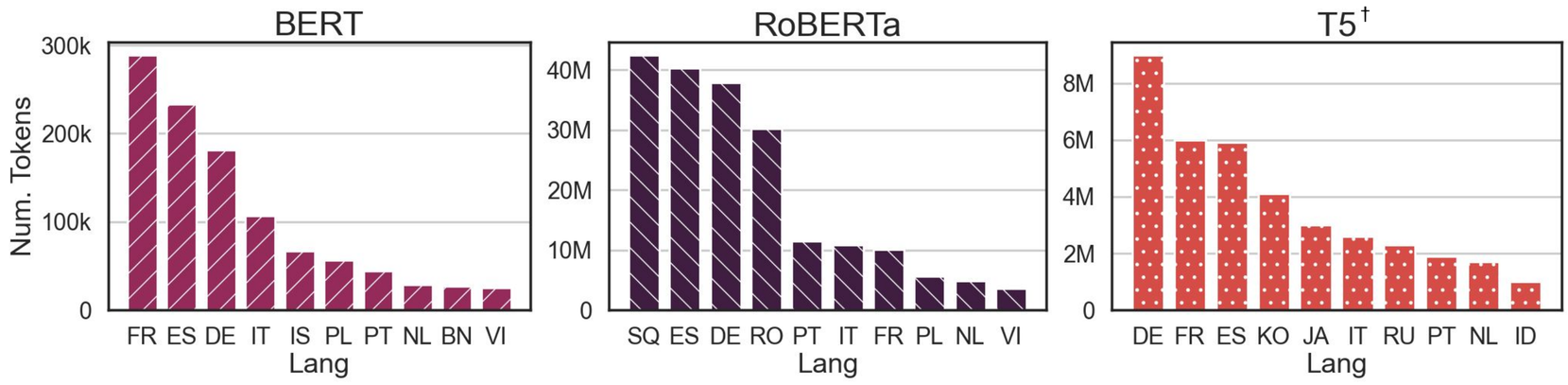
- ➔ All English corpora analyzed contain non-EN text
- ➔ Model's ability to transfer across languages is correlated with data leakage

(1) How Much Non-English Text is In English Pretraining Data?

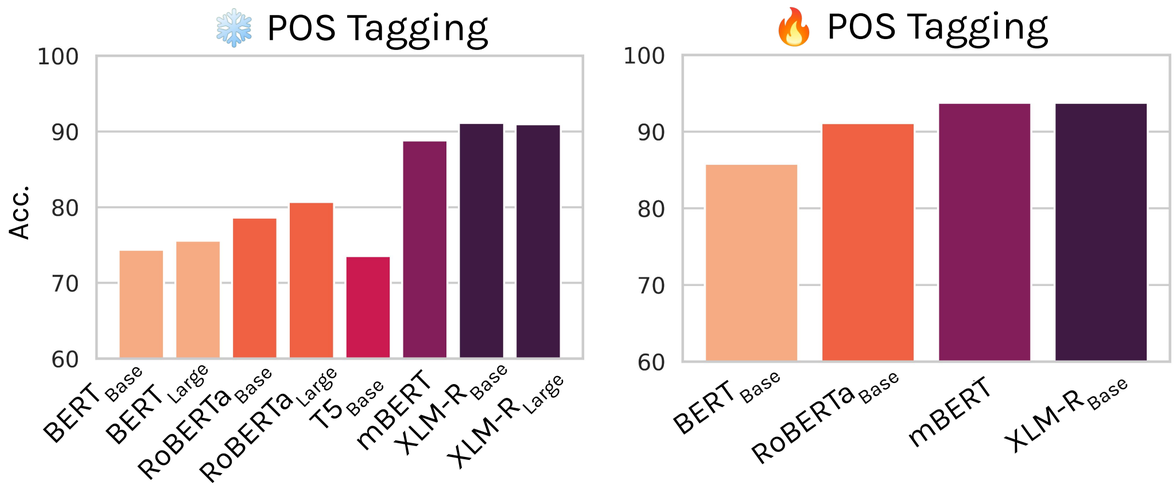


Total Amount of Non-English Text in Pretraining Corpora: small %s but large absolute quantities of tokens

Top 10 non-EN languages in Each Model



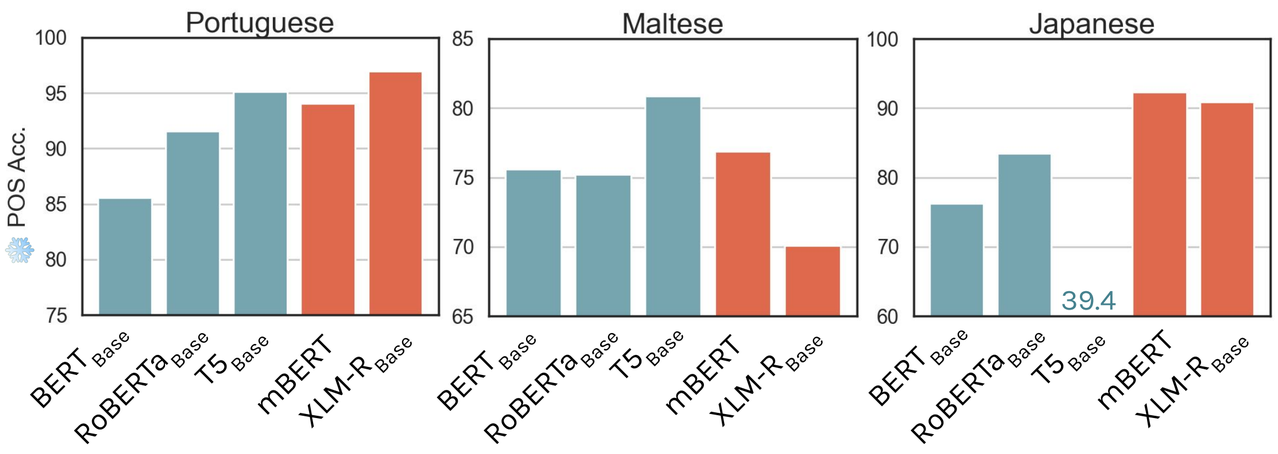
(2) Transferring English Models To Other Languages



Average performance of models across 47 non-EN languages

- ➔ RoBERTa > other EN models
- ➔ Finetuning shrinks gap between EN and multilingual LMs

Performance by **English** and **multilingual** models on: high-resource (PT), low-resource (MT), and non-Latin script (JA)

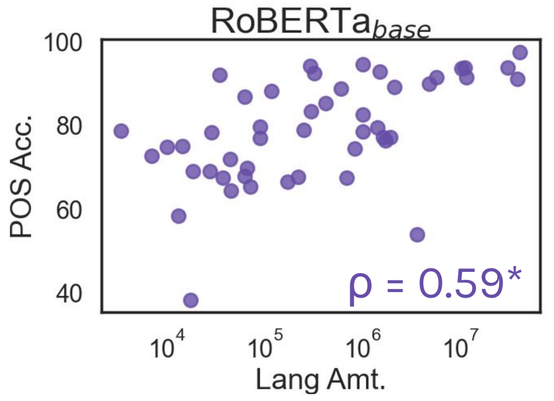


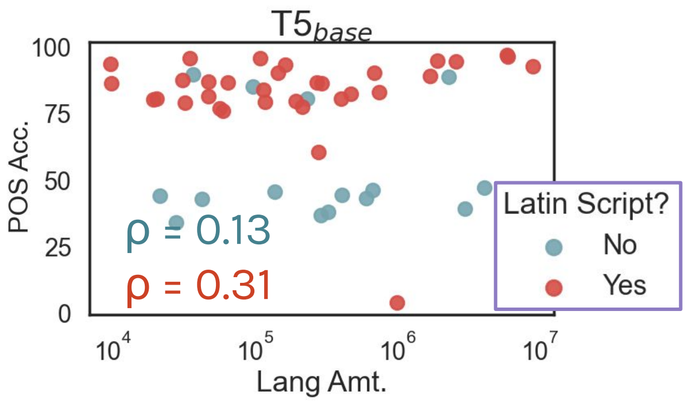
What Types of Non-English Text?

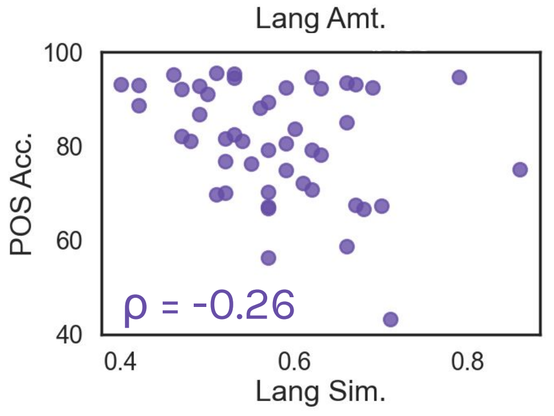
Examples of Data Leakage	Non-English	**Moraliska argument utgår ifrån våra moraliska intuitioner…** (OpenWebText)
	Bilingual	The German blazon reads: "**Von Silber über Schwarz geteilt…**" (Wikipedia)
	Trans-lation	**Εκείνη δεν μπορούσε να πληρώσει** [She couldn't pay.] (BookCorpus)
	Entities	2012 Playhouse Presents **ウィルシリーズ1、エピソード1**: "The Minor Character" (C4)
Language Classifier Errors	English	**"Dere's buzzards circlin' ova dem trees."** (BookCorpus)
	Noise	**M D | X O X | O O O = A** (Wikipedia)

(3) Potential Reasons for Cross-Lingual Transfer

- ➔ Amount (**Amt.**) of non-EN data in pretraining corpus?
- ➔ Similarity (**Sim.**) of target lang to EN?







blvns.github.io
@terrablvns

Check out the paper!